View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Bournemouth University Research Online

The 2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), DOI: 10.1109/EAIS.2015.7368803

A Non-parametric Hierarchical Clustering Model

Saad Mohamad, Abdelhamid Bouchachia, *Senior Member, IEEE,* and Moamar Sayed-Mouchaweh

***Abstract*—We present a novel non-parametric clustering model using Gaussian mixture model (NHCM). NHCM uses a novel Dirichlet process (DP) prior allowing for more flexible modeling of the data, where the base distribution of DP is itself an infinite mixture of Gaussian conjugate prior. NHCM can be thought of as hierarchical clustering model, in which the low level base prior governs the distribution of the data points forming sub-clusters, and the higher level prior governs the distribution of the sub-clusters forming clusters. Using this hierarchical configuration, we can maintain low complexity of the model and allow for clustering skewed complex data. To perform inference, we propose a Gibbs sampling algorithm. Empirical investigations have been carried out to analyse the efficiency of the proposed clustering model.**

I. Introduction

FLat clustering creates a flat set of clusters without any explicit structure that would relate clusters to each other. While hierarchical clustering aims at producing a hierarchical series of nested clusters, ranging from clusters of individual points at the bottom to an all-inclusive cluster at the top. However, real world data often exhibits a hierarchical structure,though sometime it is implicit. Thus, many levels of grouping can be done. In this paper, we propose a two-level clustering algorithm. Unlike the hierarchical Dirichlet process prior [1], different clusters in the same level do not share components allowing for modeling of multi-center distributions. The proposed NHCM model uses a novel Bayesian non-parametric prior. This prior can be simply described as a distribution on the space of distributions on distributions. Its low level is a distribution over distribution, which itself is distributed according to the higher level distribution.

Probabilistic machine learning assumes underlying distributions for the observed data. To model these distributions, in the literature,we encounter two approaches; optimization and Bayesian approaches. The difference between the optimization and the Bayesian modeling is that the former uses a point estimate of the model parameter as in GGMM [2], while the Bayesian approach models the uncertainty of the model parameters using a distribution. We also distinguish parametric and non-parametric approaches. For parametric approaches, data is represented by models using a fixed and finite number of parameters. Thus, an assumed number of parameters has to be determined apriori; like the number of clusters in a parametric clustering. In contrast, in non-parametric modeling, the number of parameters can grow with the sample size. In addition, by allowing the complexity of the model to be unbounded, under-fitting gets mitigated. Over-fitting, on the other hand, is mitigated using a Bayesian approach. For more comparison between the two modeling approaches, please refer to [3]. In the following, we will briefly explain dirichlet and hierarchical dirichlet process. For more details, please refer to [1], [3].

S. Mohamad and A. Bouchachia are with the Department of Computing, Bournemouth University, Poole, UK. e-mail:{SMohamad, abouchachia}@bournemouth.ac.uk

M. Sayed-Mouchaweh is with Ecole des Mines, Douai, France. e-mail:moamar.sayed-mouchaweh@mines-douai.fr

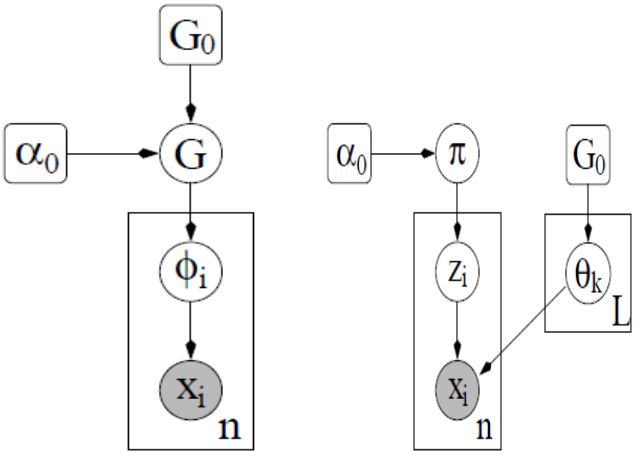
A. Background

1) Dirichlet process: DP is one of the most popular prior used in Bayesian non-parametric model. Its first use by the machine learning community dates back to [4], [5]. In general, stochastic process is probability distribution over a space of paths which describe the evolution of some random value over time. DP is a family of stochastic processes whose paths are probability distributions. It can be seen as an infinite-dimensional generalization of Dirichlet distribution, where it is a prior over the space of countably infinite distributions. In the literature, DP has been constructed with different ways, the most well-known constructions are: infinite mixture model [5], distribution over distribution [6], Polya-urn scheme [7] and stick-breaking [8]. For more details, interested reader is referred to [3].

Figure 1 shows two DP representations, the graphical model and the infinite mixture model with number of clusters L goes to ∞. Infinite mixture model is simply a generalization of the finite mixture model, where DP prior with infinite parameters is used instead of Dirichlet distribution prior with fixed number of parameters. The finite mixture model can be represented by the following equations:

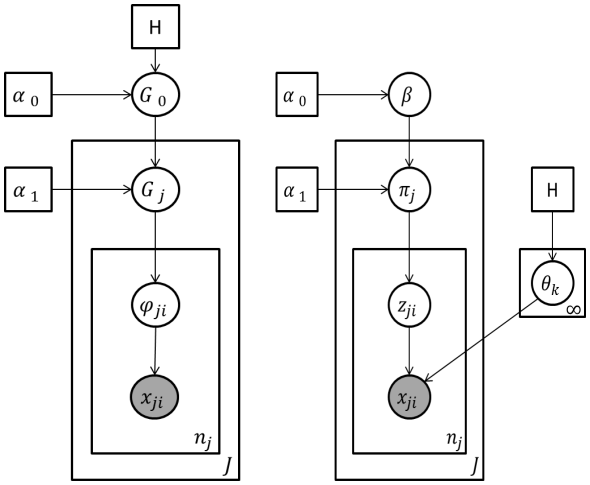
$$\begin{aligned} \boldsymbol{\pi}|\alpha_0 &\sim Dirichlet(\alpha_0/L, ..., \alpha_0/L) \\ z_i|\boldsymbol{\pi} &\sim Discrete(\pi_1, ..., \pi_L) \\ \boldsymbol{\theta_k}|G_0 &\sim G_0 \\ \boldsymbol{x_i}|z_i, \boldsymbol{\theta} &\sim F(\boldsymbol{\theta_{z_i}}) \end{aligned} \tag{1}$$

$F(\boldsymbol{\theta_{z_i}})$ denotes the distribution of the observation $\boldsymbol{x_i}$ given $\boldsymbol{\theta_{zi}}$, where $\boldsymbol{\theta_{zi}}$ is the parameter vector associated with component z_i. Here z_i indicates which latent cluster is associated with observation $\boldsymbol{x_i}$. Indicator z_i is drawn from a Discrete distribution governed by parameter $\boldsymbol{\pi}$ drawn from dirichlet distribution parametrized by α_0. We can simply say that $\boldsymbol{x_i}$ is distributed according to a mixture component drawn from $\boldsymbol{\theta_k}$ prior distribution G_0 and picked with probability given by the vector of mixing proportions $\boldsymbol{\pi}$. The model represented by Eq.(1) above is a finite mixture model, where L is the fixed number of parameters (components). The infinite mixture model can be derived by letting $L \to \infty$, then $\boldsymbol{\pi}$ can be represented as an infinite mixing proportion distributed



(a) Graphical model (b) Finite mixture model

Figure. 1 DP mixture model representations



(a) Graphical model (b) Finite hierarchical multiple mixture model

Figure. 2 Hirarchical multi-layer Dirichlet process mixture model

according to stick-breaking distribution $GEM(\alpha)$ [8]. Thus, Eq.(1) can be equivalently expressed according to the graphical representation as:

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \boldsymbol{\theta_i}|G &\sim G \\ \boldsymbol{x_i}|\boldsymbol{\theta_i} &\sim F(\boldsymbol{\theta_i}) \end{aligned} \tag{2}$$

where $G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta_i}}$ is drawn from DP prior, $\delta_{\boldsymbol{\theta_i}}$ is a Dirac delta function centred at $\boldsymbol{\theta_i}$. Technically, DP is a distribution over distribution [6], where $DP(G_0, \alpha)$, is parametrized by the base distribution G_0, and the concentration parameter α. Since DP is distribution over distributions, a draw G from it is a distribution. Thus, we can sample $\boldsymbol{\theta_i}$ from G. Back to Eq.(1), by integrating over the mixing proportion $\boldsymbol{\pi}$, we can write the prior for z_i as conditional probability of the following form [9]:

$$p(z_i = c|z_1, ..., z_{i-1}) = \frac{n_c^{-i} + \alpha_0/L}{i - 1 + \alpha_0} \tag{3}$$

where n_c^{-i} is the number of z_i for $j < i$ that are equal to c. by letting L goes to infinity we get the following equations:

$$\begin{aligned} P(z_i = c|z_1, ..., z_{i-1}) &\rightarrow \frac{n_c^{-i}}{i - 1 + \alpha_0} \\ P(z_i \neq z_j \, for \, all \, j < i|z_1, ..., z_{i-1}) &\rightarrow \frac{\alpha_0}{i - 1 + \alpha_0} \end{aligned} \tag{4}$$

For an observation $\boldsymbol{x_i}$ with $z_i \neq z_j \, for \, all \, j < i$, a new component gets created with indicator $z_i = c_{new}$. For more details about the process of obtaining the prior distribution, reader is referred to [9].

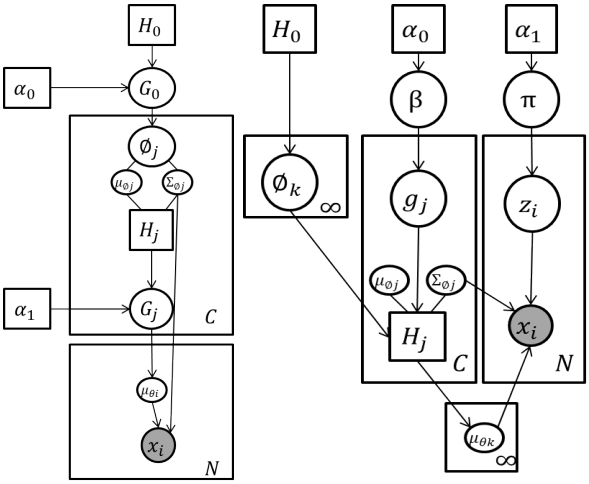
2) Hierarchical Dirichlet process: Hierarchical Dirichlet process (HDP) prior [1] has been proposed to link mixture models sharing same characteristics. Figures 2 shows the graphical model and its equivalent representation as HDP mixture model in terms of the stick-breaking representation. The equations of the graphical model are as follows:

$$\begin{aligned} G_0|H, \alpha_0 &\sim DP(\alpha_0, H) \\ G_j|G_0, \alpha_1 &\sim DP(\alpha_1, G_0) \\ \boldsymbol{\theta_{ji}}|G_j &\sim G_j \\ \boldsymbol{x_{ji}}|\boldsymbol{\theta_{ji}} &\sim F_1(\boldsymbol{\theta_{ji}}) \end{aligned} \tag{5}$$

Here, it is assumed that there are J groups of data, where the j^{th} group is denoted as $(\boldsymbol{x_{ji}})_{\boldsymbol{i=1}}^{\boldsymbol{n_j}}$. Each group has a set of parameters $(\boldsymbol{\theta_{ji}})_{\boldsymbol{i=1}}^{\boldsymbol{n_j}}$ governing the distribution $F(\boldsymbol{\theta_{ji}})$ of data. As shown in Eq. (5), G_0 representing the base distribution of low-level mixture of DP prior, is itself drawn from the DP prior. Thus, each draw G_j from DP prior $DP(\alpha_1, G_0)$ has a shared discrete base distribution G_0 allowing the different distributions to share the same set of atoms but have distinct sets of weights. One can think of G_j as a distribution on distributions and the whole model as a mixture of distribution on distributions. Thus, it is can be used to allow some characteristics to be shared between two similar problems.

II. Proposed approach

Unlike the hierarchical approach, which has one continuous base distribution of the higher level DP prior, we propose a novel DP prior with multi-base continuous distributions. Instead of having one Gaussian conjugate prior as base distribution for the higher-level DP prior, an infinite mixture of samples from a DP prior sample imposes the parameters of the base distribution of the model DP prior. The model can be thought of as a multi-layer clustering model, where the high-layer clusters govern the low-layer sub-clusters parameters distribution. Figure 3 shows graphical representation of the model. The nodes in Fig 3 correspond to the variables of our model. The arrows correspond to dependencies between the variables, and the lines are just to show the sub-variables of



(a) Graphical model (b) Infinite multi-base DP mixture model

Figure. 3 Representaions of NHCM

the main variable. The generative model shown in Fig.3a is represented as follows:

$$\begin{aligned} H_0|(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}), v_0, k_0 &\equiv NIW(.|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}, k_0, v_0) \\ G_0|\alpha_0, H_0 &\sim DP(\alpha_0, H_0) \\ \boldsymbol{\phi_j} = (\boldsymbol{\mu_{\phi j}}, \boldsymbol{\Sigma_{\phi j}})|G_0 &\sim G_0 \\ H_j|(\boldsymbol{\mu_{\phi j}}, \boldsymbol{\Sigma_{\phi j}}), q &\equiv N(.|\boldsymbol{\mu_{\phi j}}, q\boldsymbol{\Sigma_{\phi j}}) \\ G_j|H_j, \alpha_1 &\sim DP(\alpha_1, H_j) \\ \boldsymbol{\mu_{\theta i}}|G_j &\sim G_j \\ \boldsymbol{x_i}|(\boldsymbol{\mu_{\theta i}}, \boldsymbol{\Sigma_{\phi j}}) &\sim N(.|\boldsymbol{\mu_{\theta i}}, \boldsymbol{\Sigma_{\phi j}}) \end{aligned} \tag{6}$$

Figure 3a, and Equation (6) show the generative process of the model. Here, we have two plates denoting replication of observations (the bottom plate) and clusters. The clusters are parametrized by the mean $\boldsymbol{\mu_{\phi j}}$ and the variance $\boldsymbol{\Sigma_{\phi j}}$ of the Normal distribution H_j. These variables are generated from G_0 drawn from DP, while H_j is the base distribution for the low level DP prior from which G_j is drawn. Here, The sub-clusters Normal distribution are parametrized by the mean $\boldsymbol{\mu_{\theta_i}}$ drawn from G_j and the variance $\boldsymbol{\Sigma_{\phi j}}$. $NIW(.|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}, k_0, v_0) = N(.|\boldsymbol{\mu_0}, k_0^{-1}\boldsymbol{\Sigma})W_{v_0}^{-1}(\boldsymbol{\Sigma}|\boldsymbol{\Sigma_0})$ denotes the Normal-Inverse-Wishart prior, where $\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}, v_0, k_0$ are the hyper-parameters of the distribution. μ_0 is the prior of the clusters' means, and Σ_0 controls the variance among their means. k_0 scales the diffusion of the clusters means, while parameter q scales the diffusion of the sub-clusters' means within cluster. v_0 is the degree of freedom of the Inverse-Wishart distribution.

We propose a Collapse Gibbs sampling algorithm to estimate the posterior of component indicators. The mathematical equations are briefly developed following the notions adopted in [9]. It is shown that applying Gibbs sampling to the model formulated in eq.(1) with mixing proportion $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ integrated out is more efficient than applying it to other posterior forms [9]. Thus, we formulate the model shown in Fig.3b as in (1):

$$\begin{aligned} \boldsymbol{\phi_k} = (\boldsymbol{\mu_{\phi k}}, \boldsymbol{\Sigma_{\phi k}}) &\sim H_0 \equiv NIW(.|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}, k_0, v_0) \\ \boldsymbol{\beta}|\alpha_0 &\sim GEM(\alpha_0) \\ g_j|\boldsymbol{\beta} &\sim Discrete(\boldsymbol{\beta}) \\ H_j|g_j, q, (\boldsymbol{\phi})_{k=1}^{\infty} &\equiv N(.|\boldsymbol{\mu_{\phi g_j}}, q\boldsymbol{\Sigma_{\phi g_j}}) \\ \boldsymbol{\mu_{\theta_k}} &\sim H_j \\ \boldsymbol{\pi}|\alpha_1 &\sim GEM(\alpha_1) \\ z_i|\boldsymbol{\pi} &\sim Discrete(\boldsymbol{\pi}) \\ \boldsymbol{x_i}|z_i, g_{z_i}, (\boldsymbol{\mu_{\theta_k}}, \boldsymbol{\Sigma_{\phi_k}})_{k=1}^{\infty} &\sim N(.|\boldsymbol{\mu_{\theta z_i}}, \boldsymbol{\Sigma_{\phi g_{z_i}}}) \end{aligned} \tag{7}$$

Similar to DP and HDP reviewed in section 1, Fig. 3b is equivalent to Fig. 3a. It is an infinite multi-base DP mixture model representation, where not only the number of mixture can grow, but also the prior governing each mixture model. In order to sample from the probability distribution of cluster indicators g_j (high-level base-prior) and sub-cluster indicators z_i (low-level base-prior) given the observations, we use Markov chain with states consist of $G = \{g_1, ..., g_C\}$ and $Z = \{z_1, ..., z_N\}$. The Collapse Gibbs sampling algorithm used is similar to [9], where we sample the sub-cluster state z_i from $z_i|\boldsymbol{z}^{-i}, \boldsymbol{g}, \boldsymbol{X}$ and cluster state g_j from the $g_j|\boldsymbol{g}^{-j}, \boldsymbol{z}, \boldsymbol{X}$

$$\begin{aligned} P(z_i = c|\boldsymbol{z^{-i}}, \boldsymbol{g}, \boldsymbol{X}) &\propto P(z_i = c, \boldsymbol{X}|\boldsymbol{z^{-i}}, \boldsymbol{g}) \\ &= P(z_i = c|\boldsymbol{z^{-i}}, \boldsymbol{g})P(\boldsymbol{X}|z_i = c, \boldsymbol{z^{-i}}, \boldsymbol{g}) \end{aligned}$$

$$\begin{aligned} P(z_i = c_{new}|\boldsymbol{z}^{-i}, \boldsymbol{g}, \boldsymbol{X}) &\propto P(z_i = c_{new}, \boldsymbol{X}|\boldsymbol{z^{-i}}, \boldsymbol{g}) \\ &= P(z_i = c_{new}|\boldsymbol{z^{-i}}, \boldsymbol{g})P(\boldsymbol{X}|z_i = c_{new}, \boldsymbol{z^{-i}}, \boldsymbol{g}) \end{aligned} \tag{8}$$

Here, c_{new} denotes new sub-cluster, $p(z_i = c|\boldsymbol{z^{-i}}, \boldsymbol{g})$ is the probability of having the new data i in sub-cluster c given the rest of the data assignments to sub-clusters $\boldsymbol{z^{-i}}$ and all the sub-cluster assignments to clusters $\boldsymbol{g}$.

$$\begin{aligned} P(z_i = c|\boldsymbol{z^{-i}}, \boldsymbol{g}) &\propto \frac{n_c^{-i}}{n_{g_c} + \alpha_1 - 1} \\ P(z_i = c_{new}|\boldsymbol{z^{-i}}, \boldsymbol{g}) &\propto \frac{\alpha_1}{n_{g_c} + \alpha_1 - 1} \end{aligned} \tag{9}$$

where n_c^{-i} is the number of data in sub-cluster c excluding x_i, and n_{g_c} is the number of data in cluster g_c. $c \in \{1, ..., C\}$, with C is the total number of sub-clusters. Equation (9) is applied for all the sub-clusters associated with the same cluster $g_{z_i} = g_c$, where the sum of the probabilities of all the possible assignments is equal to one. Because of the use of conjugate prior for the Gaussian parameters distribution, a close form solution for the likelihood $P(\boldsymbol{X}|z_i = c, \boldsymbol{z^{-i}}, \boldsymbol{g})$ and $P(\boldsymbol{X}|z_i = c_{new}, \boldsymbol{z^{-i}}, \boldsymbol{g})$ can be obtained. Similar to the simple DP model, we will end up with the Student's t-distribution. The mathematical solution and the parameters of the distribution are detailed as follows:

$$\begin{aligned} &P(\boldsymbol{X}|z_i = c, \boldsymbol{z^{-i}}, \boldsymbol{g}) = \\ &P(\boldsymbol{x_i}, \boldsymbol{X}_{\boldsymbol{z_i}}^{-\boldsymbol{i}}, \boldsymbol{X}_{\boldsymbol{g_{z_i}}}^{-\boldsymbol{z_i}}, \boldsymbol{X}^{-\boldsymbol{g_{z_i}}}|z_i = c, \boldsymbol{z^{-i}}, \boldsymbol{g}) \end{aligned} \tag{10}$$

where $\boldsymbol{X}_{\boldsymbol{z_i}}^{-\boldsymbol{i}}, \boldsymbol{X}_{\boldsymbol{g_{z_i}}}^{-\boldsymbol{z_i}}, \boldsymbol{X}^{-\boldsymbol{g_{z_i}}}$ are respectively the data in sub-cluster z_i excluding $\boldsymbol{x_i}$, data in cluster $\boldsymbol{g_{z_i}}$ excluding z_i, and

the rest of data. Given that $\boldsymbol{g_{z_i}} = g_c$, eq.(10) can be expressed as follows:

$$P(\boldsymbol{x_i}, \boldsymbol{X_{z_i}^{-i}}, \boldsymbol{X_{g_{z_i}}^{-z_i}}, \boldsymbol{X^{-g_{z_i}}} | z_i = c, \boldsymbol{z^{-i}}, \boldsymbol{g}) \propto P(\boldsymbol{x_i} | \boldsymbol{X_c^{-i}}, \boldsymbol{X_{g_c}^{-c}}) \tag{11}$$

By plugging in the sub-cluster parameter $(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}})$, we get the following equation:

$$P(\boldsymbol{x_i} | \boldsymbol{X_c^{-i}}, \boldsymbol{X_{g_c}^{-c}}) = \int_{\boldsymbol{\Sigma_{\phi g_c}}} \int_{\boldsymbol{\mu_{\theta c}}} P(\boldsymbol{x_i} | \boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_c^{-i}}, \boldsymbol{X_{g_c}^{-c}}) d_{\boldsymbol{\mu_{\theta c}}} d_{\boldsymbol{\Sigma_{\phi g_c}}} \tag{12}$$

where $x_i | \boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}} \sim N(.|\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}})$ and $(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}}) | \boldsymbol{X_c^{-i}}, \boldsymbol{X_{g_c}^{-c}} \sim NIW(.|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{k}, \hat{v})$. Hence, by integrating out $(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}})$, we obtain the student's t-distribution as already mentioned above. In the following, we will compute the parameters of the Normal-Inverse-Wishart prior $NIW(.|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{k}, \hat{v})$, then the parameters of the student's t-distribution $t_v(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

$$P(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_c^{-i}}, \boldsymbol{X_{g_c}^{-c}}) \propto P(\boldsymbol{X_c^{-i}} | \boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}) \tag{13}$$

$$P(\boldsymbol{\mu_{\theta c}}, \boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}) = P(\boldsymbol{\mu_{\theta c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}) \tag{14}$$

$$P(\boldsymbol{\mu_{\theta c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}}) = \int_{\boldsymbol{\mu_{\phi g_c}}} P(\boldsymbol{\mu_{\theta c}} | \boldsymbol{\mu_{\phi g_c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}}) d_{\boldsymbol{\mu_{\phi g_c}}} \tag{15}$$

$$\begin{aligned} P(\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}}) &\propto \int_{\boldsymbol{\mu_{\theta t_1}}} P(\boldsymbol{X_{t_1}} | \boldsymbol{\mu_{\theta t_1}}, \boldsymbol{\Sigma_{\phi g_c}}) \\ &P(\boldsymbol{\mu_{\theta t_1}} | \boldsymbol{\mu_{\phi g_c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{X_{g_c}^{-(c,t_1)}}, \boldsymbol{\Sigma_{\phi g_c}}) d_{\boldsymbol{\mu_{\theta t_1}}} \\ &\vdots \\ P(\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{X_{g_c}^{-(c,..,t_{n-1})}}, \boldsymbol{\Sigma_{\phi g_c}}) &\propto \int_{\boldsymbol{\mu_{\theta t_n}}} P(\boldsymbol{X_{t_n}} | \boldsymbol{\mu_{\theta t_n}}, \boldsymbol{\Sigma_{\phi g_c}}) \\ &P(\boldsymbol{\mu_{\theta t_n}} | \boldsymbol{\mu_{\phi g_c}}, \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{\Sigma_{\phi g_c}}) d_{\boldsymbol{\mu_{\theta t_n}}} \end{aligned} \tag{16}$$

where $t_1, ..., t_n$ are all the sub-clusters in cluster g_c excluding c, that is $g_{t_i} = g_c$ for $i \in \{1, ..., n\}$. Equation (16) can be solved analytically as all the random variable are normally distributed. To compute the parameter of $\boldsymbol{\mu_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}} \sim N(.|\dot{\boldsymbol{\mu}}, \dot{\boldsymbol{\Sigma}})$, we roll up eq.(16) by starting from t_n. The final solution can be expressed as follows:

$$\begin{aligned} \dot{\boldsymbol{\mu}} &= \Big(k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}}\Big)^{-1} \Big(k_0 * \boldsymbol{u_0} + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}} \bar{\boldsymbol{x}}_{\boldsymbol{t_i}}\Big) \\ \dot{\boldsymbol{\Sigma}} &= \Big(k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}}\Big)^{-1} \boldsymbol{\Sigma_{\phi g_c}} \end{aligned} \tag{17}$$

where n_{t_i} is the number of data in sub-cluster t_i, and $\bar{\boldsymbol{x}}_{\boldsymbol{t_i}}$ is the mean of the data in t_i. Now, we can compute the parameter of $\boldsymbol{\mu_{\theta c}} | \boldsymbol{X_{g_c}^{-c}}, \boldsymbol{\Sigma_{\phi g_c}} \sim N(.|\ddot{\boldsymbol{\mu}}, \ddot{\boldsymbol{\Sigma}})$ in equation (15):

$$\begin{aligned} \ddot{\boldsymbol{\mu}} &= \Big(k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}}\Big)^{-1} \Big(k_0 * \boldsymbol{u_0} + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}} \bar{\boldsymbol{x}}_{\boldsymbol{t_i}}\Big) \\ \ddot{\boldsymbol{\Sigma}} &= q\boldsymbol{\Sigma_{\phi g_c}} + \dot{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma_{\phi g_c}} \Big(q + (k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}})^{-1}\Big) \end{aligned} \tag{18}$$

We still need to compute the parameter of $\boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}} \sim W_{\bar{v}}^{-1}(\bar{\boldsymbol{\Sigma}})$ in order to find the Normal-Inverse-Wishart parameters in (12).

$$\begin{aligned} P(\boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c}}) &\propto P(\boldsymbol{X_{t_1}} | \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-c,t_1}}) \\ &\vdots \\ P(\boldsymbol{\Sigma_{\phi g_c}} | \boldsymbol{X_{g_c}^{-(c,..,t_{n-1})}}) &\propto P(\boldsymbol{X_{t_n}} | \boldsymbol{\Sigma_{\phi g_c}}) P(\boldsymbol{\Sigma_{\phi g_c}}) \end{aligned} \tag{19}$$

Similar to eq.(16), we roll up by starting from the bottom.

$$\begin{aligned} \bar{v} &= v_0 + \sum_{i=1}^{n} n_{t_i} \\ \bar{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma_0} + \sum_{i=1}^{n} \boldsymbol{S_{t_i}} \end{aligned} \tag{20}$$

where $\boldsymbol{S_{t_i}}$ is the covariance of the data in sub-cluster t_i. Finally, we are able to compute the posterior distribution of the Normal-Inverse-Wishart in eq.(13):

$$\begin{aligned} \ddot{k}^{-1} &= \Big(q + (k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}})^{-1}\Big) \\ \hat{\boldsymbol{\mu}} &= \frac{\ddot{k}}{\ddot{k} + n_c - 1} \ddot{\boldsymbol{\mu}} + \frac{n_c - 1}{\ddot{k} + n_c - 1} \bar{\boldsymbol{x}}_{\boldsymbol{c}}^{-\boldsymbol{i}} \\ \hat{\boldsymbol{\Sigma}} &= \bar{\boldsymbol{\Sigma}} + \boldsymbol{S_c^{-i}} + \frac{\ddot{k}(n_c - 1)}{n_c - 1 + \ddot{k}} (\bar{\boldsymbol{x}}_{\boldsymbol{c}}^{-\boldsymbol{i}} - \ddot{\boldsymbol{\mu}})(\bar{\boldsymbol{x}}_{\boldsymbol{c}}^{-\boldsymbol{i}} - \ddot{\boldsymbol{\mu}})^T \\ \hat{v} &= \bar{v} + n_c - 1 \\ \hat{k} &= \ddot{k} + n_c - 1 \end{aligned} \tag{21}$$

Now, the integral in (12) leads to the student's t-distribution with the following parameters

$$\begin{aligned} \boldsymbol{\mu} &= \hat{\boldsymbol{\mu}} \\ \boldsymbol{\Sigma} &= \frac{\hat{\boldsymbol{\Sigma}}(\hat{k} + 1)}{\hat{k}(\hat{v} - d + 1)} \\ v &= \hat{v} - d + 1 \end{aligned} \tag{22}$$

where d is the data dimension. As for $P(z_i = c_{new} | \boldsymbol{z^{-i}}, \boldsymbol{g}, \boldsymbol{X})$, we get the following parameters by setting all the numbers of data in sub-clusters to zeros.

$$\begin{aligned} \boldsymbol{\mu'} &= \boldsymbol{\mu_0} \\ \boldsymbol{\Sigma'} &= \frac{\boldsymbol{\Sigma_0}\big(k_0(q * k_0 + 1)^{-1} + 1\big)}{k_0(q * k_0 + 1)^{-1}(v_0 - d + 1)} \\ v &= v_0 - d + 1 \end{aligned} \tag{23}$$

Therefore, Equation (8) can be expressed as follows:

$$P(z_i = c|\boldsymbol{z^{-i}}, \boldsymbol{g}, \boldsymbol{X}) \propto \frac{n_c^{-i}}{n_{g_c} + \alpha_1 - 1} t_v(\boldsymbol{x_i}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$P(z_i = c_{new}|\boldsymbol{z^{-i}}, \boldsymbol{g}, \boldsymbol{X}) \propto \frac{\alpha_1}{N + \alpha_1 - 1} t_{v'}(\boldsymbol{x_i}|\boldsymbol{\mu'}, \boldsymbol{\Sigma'}) \tag{24}$$

Similarly, the conditional probability of sub-cluster assignments to clusters given the data assignments to sub-clusters can be written as follow:

$$\begin{aligned} P(g_j = r|\boldsymbol{g^{-j}}, \boldsymbol{z}, \boldsymbol{X}) &\propto P(g_j = r, \boldsymbol{X}|\boldsymbol{g^{-j}}, \boldsymbol{z}) \\ &= P(g_j = r|\boldsymbol{g^{-j}}, \boldsymbol{z}) P(\boldsymbol{X}|g_j = r, \boldsymbol{g^{-j}}, \boldsymbol{z}) \end{aligned}$$

$$\begin{aligned} P(g_j = r_{new}|\boldsymbol{g}^{-j}, \boldsymbol{z}, \boldsymbol{X}) &\propto P(g_j = r_{new}, \boldsymbol{X}|\boldsymbol{g^{-j}}, \boldsymbol{z}) \\ &= P(g_j = r_{new}|\boldsymbol{g^{-j}}, \boldsymbol{z}) P(\boldsymbol{X}|g_j = r_{new}, \boldsymbol{g^{-j}}, \boldsymbol{z}) \end{aligned} \tag{25}$$

where g_{new} denotes new cluster, and $r \in \{1, ..., R\}$ with R is the total number of sub-cluster assignments to the existing clusters. Here, j refers to the sub-clusters $j \in \{1, ..., C\}$.

$$P(g_j = r|\boldsymbol{g^{-j}}, \boldsymbol{z}) \propto N_r^{-j}$$
$$P(g_j = r_{new}|\boldsymbol{g^{-j}}, \boldsymbol{z}) \propto \alpha_0 \tag{26}$$

N_r is the number of sub-clusters in cluster r excluding j. Likewise the data assignments, we will end up with student's t-distribution. The mathematical solution and the parameters of the distribution are detailed as follows:

$$\begin{aligned} P(\boldsymbol{X}|z_i = c, \boldsymbol{g^{-j}}, \boldsymbol{z}) &\propto P(\boldsymbol{X_j}, \boldsymbol{X_{g_j}^{-j}}|g_j = r, \boldsymbol{g^{-j}}, \boldsymbol{z}) \\ &= P(\boldsymbol{X}_j, \boldsymbol{X_r^{-j}}) \end{aligned} \tag{27}$$

where $\boldsymbol{X_j}, \boldsymbol{X_r^{-j}}$ are respectively the data in sub-cluster j, and the sub-clusters in cluster r excluding j. By plugging in the sub-cluster parameter $(\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}})$, we get the following equation:

$$\begin{aligned} P(\boldsymbol{X_j}|\boldsymbol{X_r^{-j}}) = \int_{\boldsymbol{\Sigma_{\phi r}}} \int_{\boldsymbol{\mu_{\theta j}}} & P(\boldsymbol{X_j}|\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}}) \\ & P(\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}}|\boldsymbol{X_r^{-j}}) d_{\boldsymbol{\mu_{\theta j}}} d_{\boldsymbol{\Sigma_{\phi r}}} \end{aligned} \tag{28}$$

Using the obtained results from the previous section, we can deduce the parameters of $(\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}})|\boldsymbol{X_r^{-j}} \sim NIW(.|\hat{\boldsymbol{\mu}}_{\mathbf{1}}, \hat{\boldsymbol{\Sigma}}_{\mathbf{1}}, \hat{k_1}, \hat{v_1})$ as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\mathbf{1}} &= \Big(k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}}\Big)^{-1} \Big(k_0 * \boldsymbol{u}_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}} \bar{\boldsymbol{x}}_{\boldsymbol{t_i}}\Big) \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{1}} &= \boldsymbol{\Sigma_0} + \sum_{i=1}^{n} \boldsymbol{S_{t_i}} \\ \hat{v}_1 &= v_0 + \sum_{i=1}^{n} n_{t_i} \\ \hat{k}_1 &= \Big(q + (k_0 + \sum_{i=1}^{n} \frac{n_{t_i}}{1 + q * n_{t_i}})^{-1}\Big)^{-1} \end{aligned} \tag{29}$$

Here, $t_1, ..., t_n$ are all the sub-clusters in cluster r excluding j, that is $g_{t_i} = r$ for $i \in \{1, ..., n\}$. As shown in fig.3, the distribution of data X_j is independent given $(\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}})$. Therefore eq.(27) can be written as follows:

$$\begin{aligned} P(\boldsymbol{X_j}|\boldsymbol{X_r^{-j}}) = \prod_{i=1}^{n_j} \int_{\boldsymbol{\Sigma_{\phi r}}} \int_{\boldsymbol{\mu_{\theta j}}} & P(\boldsymbol{x_{j,i}}|\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}}) \\ & P(\boldsymbol{\mu_{\theta j}}, \boldsymbol{\Sigma_{\phi r}}|\boldsymbol{X_r^{-j}}) d_{\boldsymbol{\mu_{\theta j}}} d_{\boldsymbol{\Sigma_{\phi r}}} \\ = \prod_{i=1}^{n_j} & t_{v_1}(x_{j,i}|\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}) \end{aligned} \tag{30}$$

where n_j is the number of data in sub-cluster j.

$$\begin{aligned} \boldsymbol{\mu_1} &= \hat{\boldsymbol{\mu}}_{\mathbf{1}} \\ \boldsymbol{\Sigma}_1 &= \frac{\hat{\boldsymbol{\Sigma}}_{\mathbf{1}}(\hat{k}_1 + 1)}{\hat{k}_1(\hat{v}_1 - d + 1)} \\ v_1 &= \hat{v}_1 - d + 1 \end{aligned} \tag{31}$$

Similarly, we have the following parameter when new cluster is created.

$$\begin{aligned} \boldsymbol{\mu'_1} &= \boldsymbol{\mu_0} \\ \boldsymbol{\Sigma'_1} &= \frac{\boldsymbol{\Sigma_0}\big(k_0(q * k_0 + 1)^{-1} + 1\big)}{k_0(q * k_0 + 1)^{-1}(v_0 - d + 1)} \\ v_1 &= v_0 - d + 1 \end{aligned} \tag{32}$$

Therefore, Equation (25) can be expressed as follows:

$$P(g_j = r|\boldsymbol{g^{-j}}, \boldsymbol{z}, \boldsymbol{X}) \propto N_r^{-j} \prod_{i=1}^{n_j} t_{v_1}(\boldsymbol{x_{j,i}}|\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$$
$$P(g_j = r_{new}|\boldsymbol{g^{-j}}, \boldsymbol{z}, \boldsymbol{X}) \propto \alpha_0 \prod_{i=1}^{n_j} t_{v'_1}(\boldsymbol{x_{j,i}}|\boldsymbol{\mu'_1}, \boldsymbol{\Sigma'_1}) \tag{33}$$

Algorithm (1) summarize NHCM. We empirically set the number of gibbs sampling iterations $ngibbs$, which is needed for reaching the stationary distribution, to 10.

III. Experiment

To discuss the efficiency of the proposed model, we analyse its behaviour in comparison with GMM with one layer DP (DPGMM). We use banana dataset for the sake of comparing NHCM against DPGMM. The banana dataset is an artificial dataset drawn from a distribution with a domain shaped in the form of banana [10]. The data is uniformly distributed along the banana trajectory, while orthogonally to the trajectory, it is normally distributed with standard deviation equal to 0.3. We generate a total of 400 points evenly divided into two classes. The banana dataset is a skewed multi-modal dataset, where clusters tend to split up to accommodate the data. Hence, using this type of dataset provides an insight into the flexibility of the algorithm. However, we believe that NHCM will show better results than classical flat clustering algorithms on any multi-modal skewed datasets. Therefore, more experiments on different synthetic and real world datasets will be carried out in the future in order to show the robustness of NHCM.

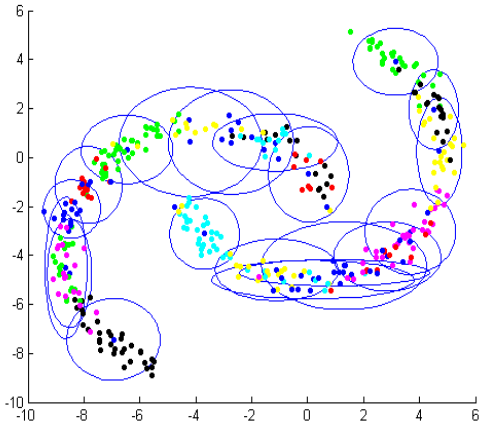
In order to allow obscure prior, we set $\alpha_1 = 1$ for NHCM and DPGMM, and $\alpha_0 = 1$ for NHCM. The mean u_0 for both

Algorithm 1 Steps of NHCM

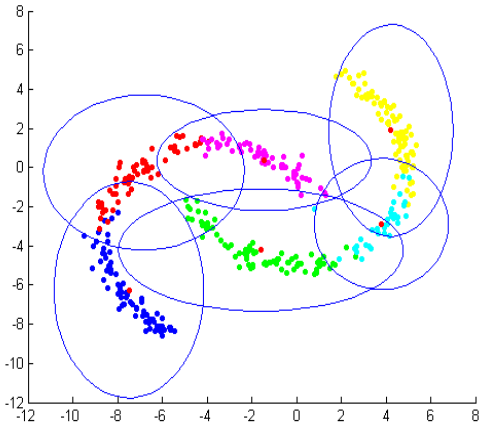
Input: data.
initialize: α_0, α_1, $\boldsymbol{\mu_0}$, $\boldsymbol{\Sigma_0}$, v_0, k_0, q to the values proposed in the experiment section. The number of clusters R to zero, and the number of sub-cluster C to zero. The number iterations before equilibrium $ngibbs$.
for all $nb = 1 : ngibbs$ **do**
  for all $i = 1 : N$ in random order **do**
    Remove $\boldsymbol{x_i}$'s sufficient statistics from its old sub-cluster z_i
    for all $c = 1 : C$ over the existing sub-clusters given that $g_c = g_{z_i}$ **do**
      Compute the probability of the $\boldsymbol{x_i}$ assignment to sub-cluster c. (Eq. (8))
    end for
    Compute the probability of the $\boldsymbol{x_i}$ assignment to new sub-cluster c_{new}. (Eq. (8))
    Sample $z_i \in p(z_i|.)$
    Add x_i to its sub-cluster z_i, and remove any empty sub-cluster.
    Remove $\boldsymbol{z_i}$'s sufficient statistics from its old cluster g_{z_i}
    for all $r = 1 : R$ over all the existing clusters **do**
      Compute the probability of the $\boldsymbol{z_i}$ assignment to cluster r. (Eq. (25))
    end for
    Compute the probability of the $\boldsymbol{z_i}$ assignment to new cluster r_{new}. (Eq. (25))
    Sample $g_{z_i} \in p(g_r|.)$
    Add z_i to its cluster g_{z_i}, and remove any empty cluster.
  end for
end for

models NHCM and DPGMM is set to be equal to the mean of the entire datasets. The degree of freedom of the Wishart distribution v_0 must be greater than the dimension of the data. We set it to $v_0 = 5$ to allows high flexibility. The rest of the parameters are empirically set as follow: $q = 3$, $k_0 = 1$ and $\Sigma_0 = I$. However, changing the parameters have slightly effect on the final results, while the convergence time changes.

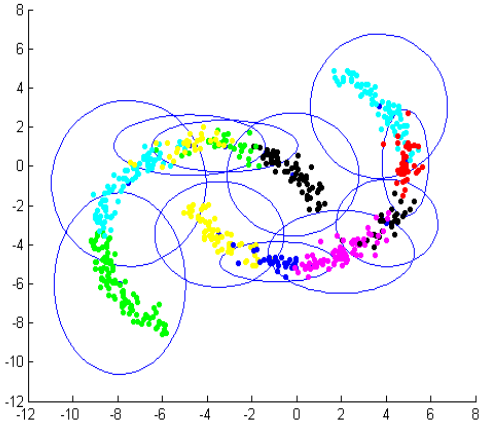
In Figure 4, the same color within the same ellipse shows that the data is in the same cluster. The ellipses are drawn according to the variance and the mean of the data in the cluster. Figures 4a shows the sub-clusters created by NHCM for banana dataset. These sub-clusters are governed by the base distribution of the clusters shown in Fig.4b. Each cluster forms a base-prior for many sub-clusters resulting in more flexible clustering. The base prior of the cluster plays the role of a distribution over the space of distributions on distributions. Figures 4c shows the results of DPGMM. It is clear that more clusters have been created. One can naively think of NHCM as a model for clustering the means of DPGMM clusters leading to higher representation with lower complexity.



(a) Low level clusters(sub-clusters) of NHCM



(b) High level clusters of NHCM



(c) Clusters of DPGMM

Figure. 4 Banana shaped dataset

IV. Conclusion

We have proposed a hierarchical clustering algorithm with unbounded complexity. The proposed model maintains low complexity by using Dirichlet process with hierarchical base

prior distribution, where close sub-clusters forms high-level clusters. It needs no prior information such as number of clusters, or merging and splitting steps. In the future, we will discard the exchangeability assumption of the data and propose a dynamic model, where inference is performed using sequential Monte Carlo. We will also add non-parametric prior over the label distribution in order to perform classification by softly gathering similar data from the same class in the same cluster. These two expansions will be finally collected to perform novelty detection using active learning.

ACKNOWLEDGMENT

A. Bouchachia was fully supported by the European Commission under the Horizon 2020 Grant 687691 related to the project: *PROTEUS: Scalable Online Machine Learning for Predictive Analytics and Real-Time Interactive Visualization*.

REFERENCES

[1] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the american statistical association*, vol. 101, no. 476, 2006.

[2] A. Bouchachia and C. Vanaret, "Gt2fc: an online growing interval type-2 self-learning fuzzy classifier," *Fuzzy Systems, IEEE Transactions on*, vol. 22, no. 4, pp. 999–1018, 2014.

[3] Y. W. Teh, "Dirichlet process," in *Encyclopedia of machine learning*. Springer, 2010, pp. 280–287.

[4] R. M. Neal, "Bayesian mixture modeling," in *Maximum Entropy and Bayesian Methods*. Springer, 1992, pp. 197–211.

[5] C. E. Rasmussen, "The infinite gaussian mixture model." in *NIPS*, vol. 12, 1999, pp. 554–560.

[6] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[7] D. Blackwell and J. B. MacQueen, "Ferguson distributions via pólya urn schemes," *The annals of statistics*, pp. 353–355, 1973.

[8] J. Sethuraman, "A constructive definition of dirichlet priors," DTIC Document, Tech. Rep., 1991.

[9] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[10] PRTOOLS. Pattern recognition tools. [Online]. Available: http://rduin.nl/prhtml/prtools/gendatb.html